II. Analysis of family resemblance and segregation analysis

- Their concepts are quite general and underlie all statistical inference.
- They are often a integral part of the other analysis.
- They do not need genetic markers to be available but can provide useful characterization of the trait under question. In this sense, it is classic yet with a modern flavour.

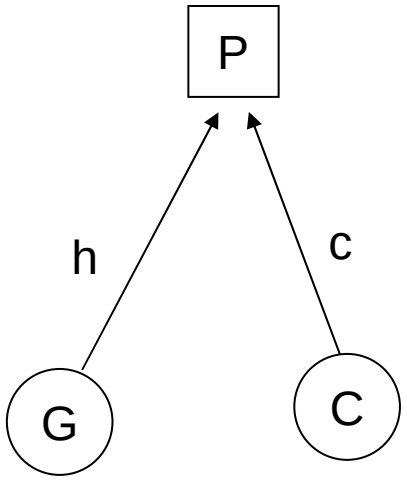
Table of contents

Assessment of familial aggregation

- For quantitative traits, **intraclass (or intrafamily) correlation coefficient (ICC)** indicates the proportion of the total variability in a phenotype that can be attributed to variability between families.
- For binary traits
  - **Prevalence** is the proportion of a population that has a disease at a given time (K)
  - Denote the prevalence of the disease in relatives of type R of affected cases (K_R)
  - **Recurrence risk ratio** ($\lambda_R = K_R/K$) can indicate power of linkage studies

Model of family resemblance

- Let P = phenotype, C = environment, E = error
- The basic model is : P = G + C + E
- Additionally, let A = additive component, D = dominance, G = A + D



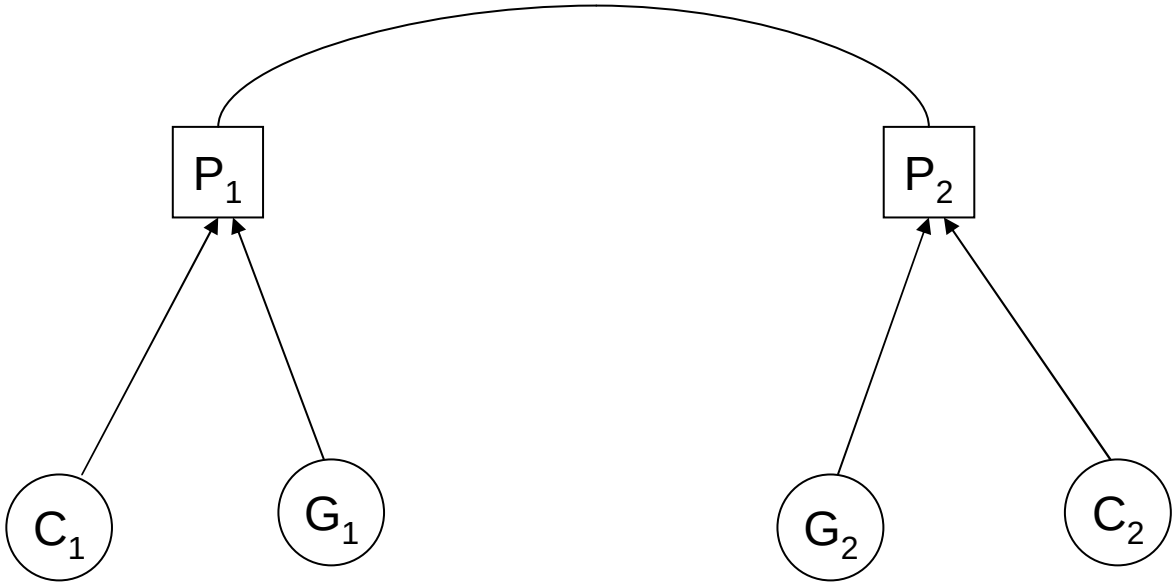
Heritability (h^2)

- For quantitative traits
  - In the narrow sense is the proportion of total phenotypic variance accounted by additive variance, V_A/V_P.
  - In the broad sense it indicates the proportion of total phenotypic variance attributable to all genetic effects including non-additive effects at individual loci and between loci, V_G/V_P.
- For binary traits, it is usually obtained through liability, an underlying, unobservable, normally-distributed trait which determines the probability that an individual develop the disease of interest.

Some clarifications about heritability

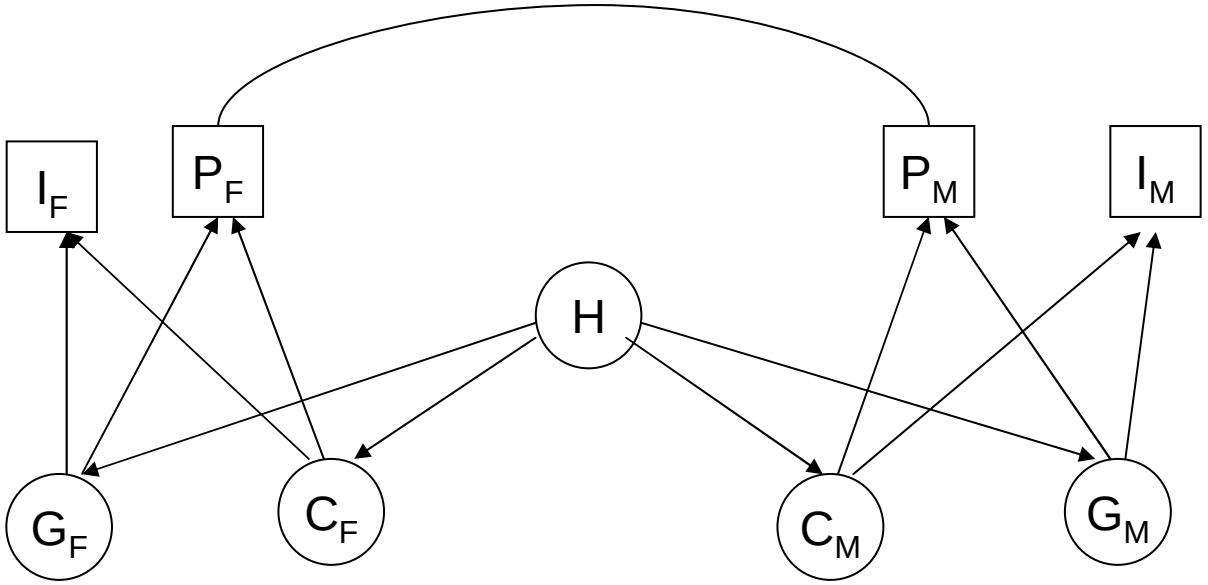
- For a binary trait, such as whether or not an individual has a disease, heritability is not the proportion of disease in the population attributable to or caused by, genetic factors.
- For a continuous trait, genetic heritability is not a measure of the proportion of an individual's score attributable to genetic factors. Heritability is not about cause *per se*, but about the causes of variation in a trait across a particular population.

Resemblance between relatives (e.g., siblings)

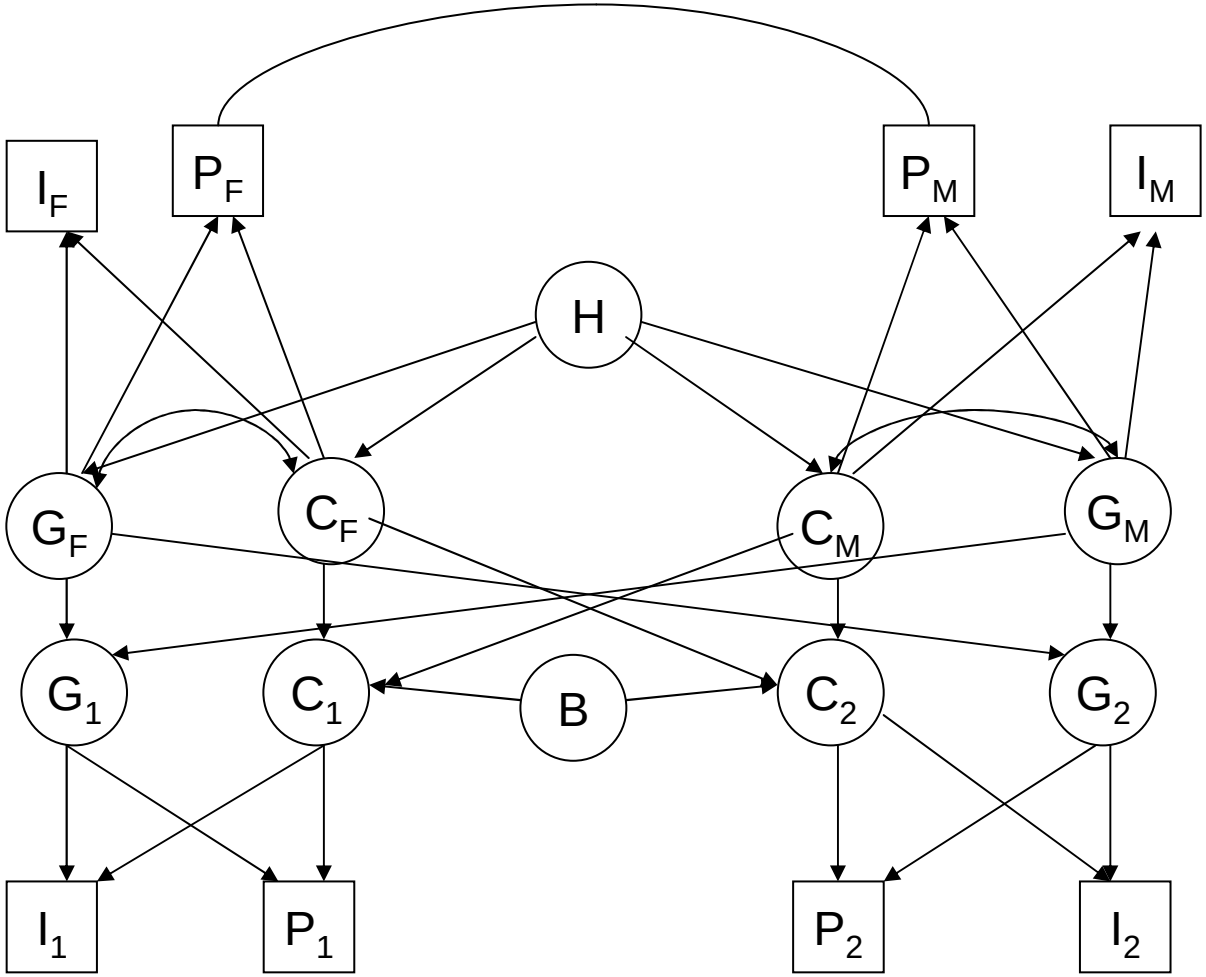


Assortative matings

- F = father, M = mother, H = homogamy, I = environmental index



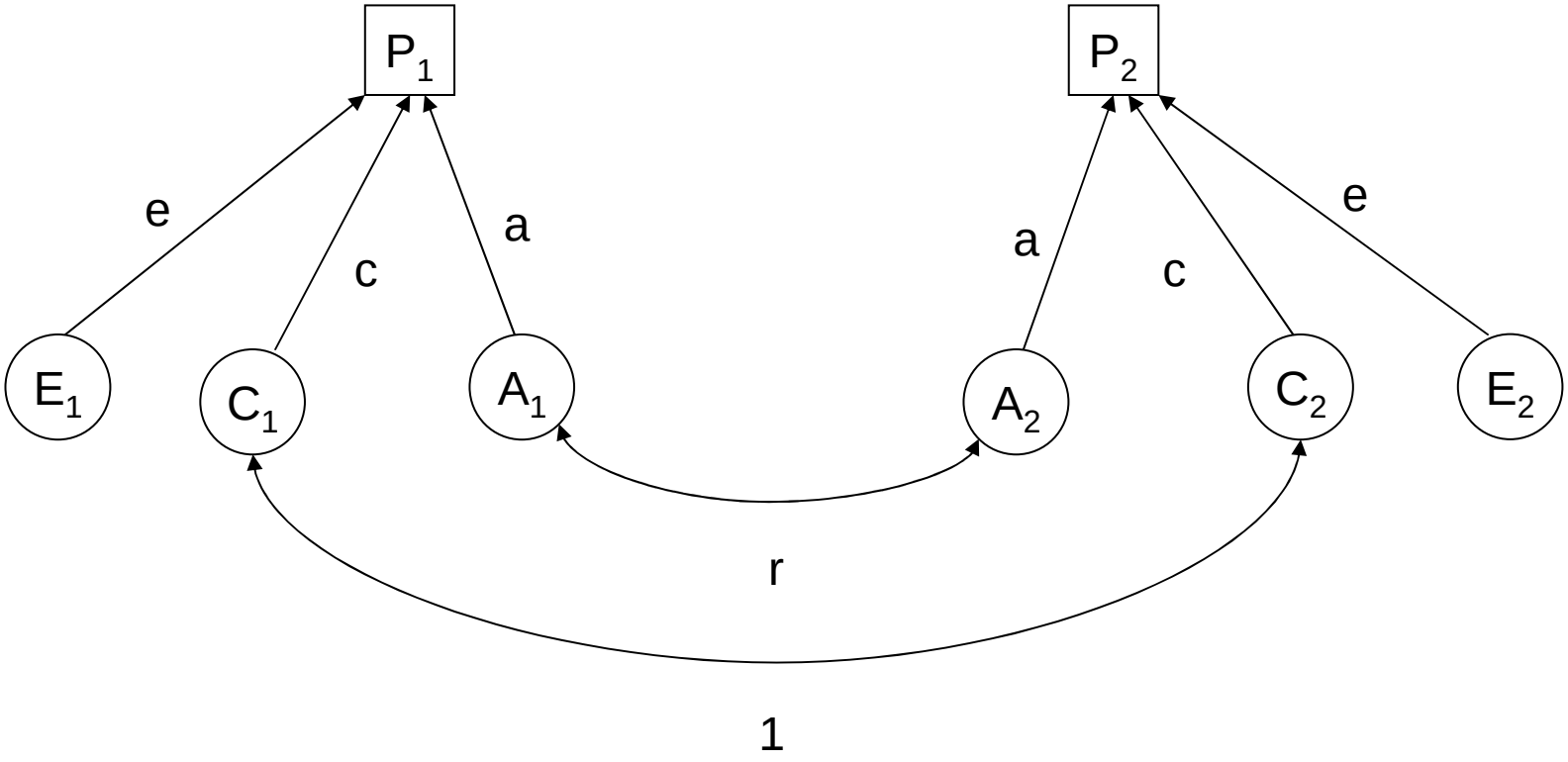
Path diagram for a nuclear family



Statistical inference

- We can assume phenotypic values of all relatives to follow multivariate normal distribution.
- The variance-covariance matrix between relatives can be obtained from path-tracing rules, expressed in terms of parameters of interest with intergenerational difference allowed.
- Parameter estimation can be achieved through maximum likelihood (GEMINI, ALMINI) directly or Fisher's transformation of correlations between relatives as implemented in the computer program PATHMIX.

Resemblance between twin pairs



- $\rho_P = r\, a^2 + c^2 + e^2$, with $r_{MZ}=1$, $r_{DZ}=0.5$
- We have $h^2=a^2=2(\rho_{MZ} - \rho_{DZ})$, $V(h^2)=4[(1- \rho_{MZ})^2/n_{MZ}+(1- \rho_{DZ})^2/n_{DZ}]$

Commingling analysis

- A mixture distribution with c components (c≥2) each with probability density function $f_i(x)$ has the density function f_M,

$$f_M(x) = \pi_1 f_1(x) + \pi_2 f_2(x) + \cdots + \pi_c f_c(x)$$

- with π_i, i=1,…,c being the mixing proportions such that $\pi_i \geq 0$, $\Sigma\pi_i = 1$
- A normal mixtures with two components is

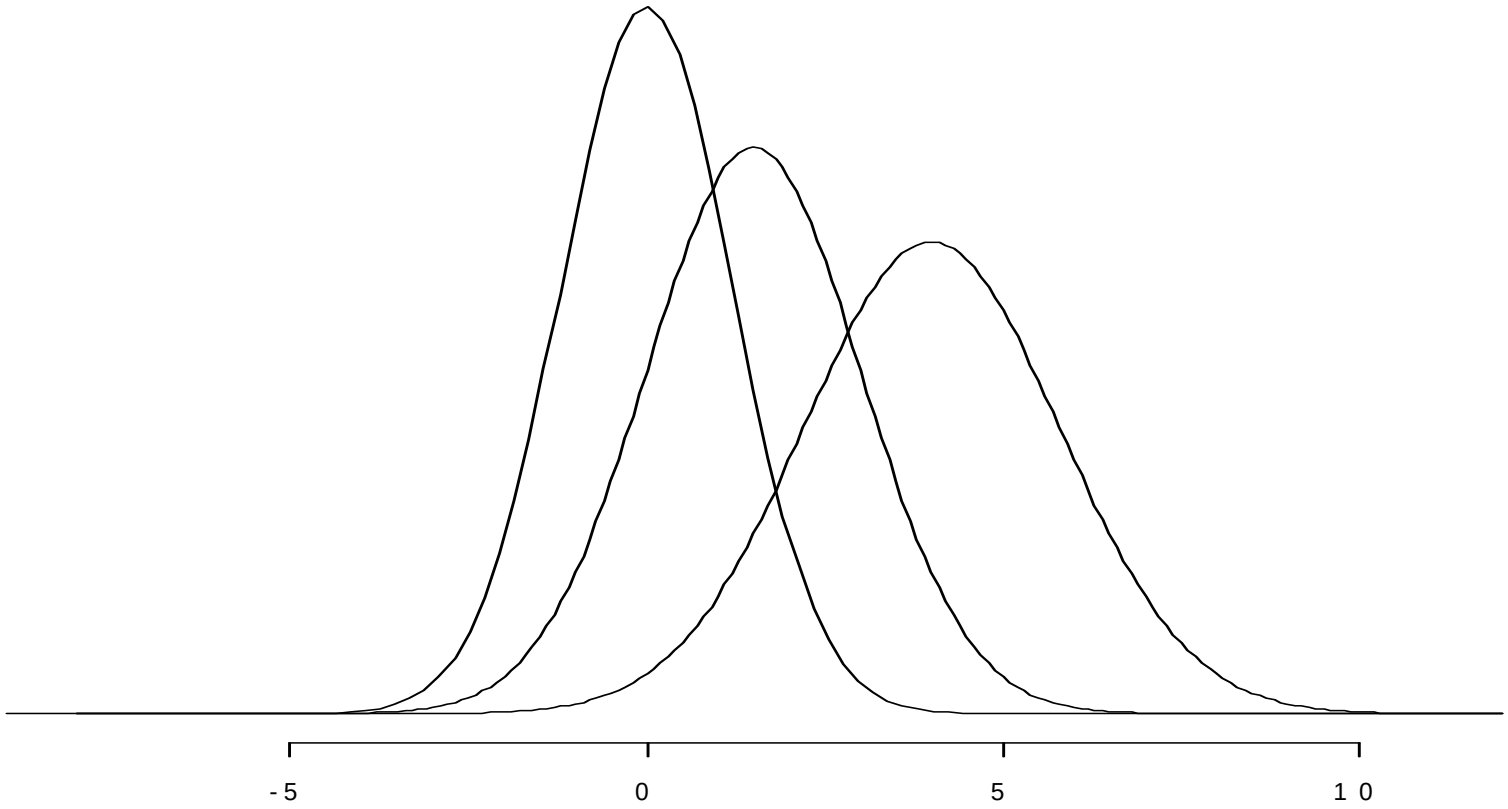
$$f_M(x) = \pi_1 \varphi(x; \mu_1, \sigma^2) + \pi_2 \varphi(x; \mu_2, \sigma^2),$$

where

$$\varphi(x; \mu, \sigma) = \frac{1}{(2\pi)^{1/2}} \exp\left[\frac{-(x-\mu)^2}{2\sigma^2}\right].$$

An example of three normal distributions

- A quantitative trait controlled by a major gene follows commingled distribution, schematic example: N(0,1.2), N(1.5, 1.5), N(4,1.8)



Commingling analysis: statistical inference

- The basic problem is to distinguish between a sample from a mixture of distributions and a sample from a single distribution
- The likelihood ratio test does not have known asymptotic distribution because the regularity condition allowing for expansion of likelihood ratio statistic does not hold.
- MacLean et al. (1976) suggested use of a (modified) Box-Cox transformation to reduce the sensitivity of the test to data from skewed distributions.
- It is of interest genetically as evidence of a genetic determination. It is a logical model for traits that are either polygenic or multifactorial. For a polygenic model, the central limit theorem dictates that the trait will be approximately normal.

Operating characteristics

(Borecki et al. Genet Epidemiol, 1994)

- Power increases with increasing sample size.
- The increase in power with increasing effect-size is a function of proportion of variance associated with the major gene effect.
- The power to reject q=0 is consistently higher for dominant vs recessive major gene effects.
- Since the types of statistical information involved in detection of major gene effects relate to commingling of genotype-specific distributions (controllable by contrasting models with same trait frequency and displacement) and segregation of alleles to offspring. It appears that the increase in information can be due to a higher proportion of informative segregating matings for dominant vs recessive traits.

Segregation analysis

- The inference of mode of inheritance from pedigree data.
- Complex segregation analysis deals with two or more distinct and functionally independent segregation parameters
  - There are two forms of a certain genetic character, neither form is rare but controlled by one (or more) loci
  - There are more than two phenotypic forms controlled by a certain genetic system (e.g., multiallelic inheritance)
  - Incomplete penetrance or different fitnesses of genotypes

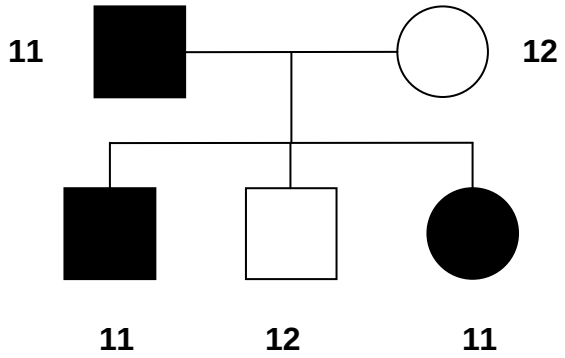
Likelihood of phenotypic data X

$$L(X) = \sum_{g_1}\sum_{g_2}\cdots\sum_{g_n}\Pi_{j=1}^{n}P(X_j|g_j)\Pi_{k=1}^{f}P(g_k)\Pi_{m=1}^{n_f}\tau(g_m|g_{m1},g_{m2})$$

- Where
  - gi = multilocus genotype of individual i
  - f = number of founders
  - nf = number of nonfounders
  - P(.) = genotypic frequency
  - P(.|.) = penetrance
  - τ(.|.) = transmission function
  - m_1,m_2 = parental indices of individual m
- τ(.|.) specifies the probabilities of the three major locus genotypes (AA, Aa, aa) transmitting the normal allele (A), to be (1,0.5,0) in the Mendelian case, while setting transmission probabilities to be the same tests the transmission of major determinants

Segregation analysis: a simple example

- Consider a simple family



- As before
  - P = allele frequencies,
  - f = penetrances,
  - *T* = transmission probabilities given parental genotypes
- The likelihood can be written as follows
- L = P(11)f(11) x P(12)(1-f(12)) x *T* (11|11,12)f(11) x *T*(12|11,12)(1-f(12)) x *T*(11|11,12)f(11)

Genetic models – modes of inheritance

- In a generalized single locus model, the disease locus has disease allele a, normal allele A, and three genotypes AA, Aa, aa, the frequency and transmission parameters can be specified in a vector.
  - Dominant: (q, 0, 1, 1)
  - Recessive: (q, 0, 1, 1)
  - Non-Mendlian: the penetrances taking any values in [0,1]
- A trait is
  - **monogenic** if it is dominated by effects of a single locus
  - **oligogenic** if it is influenced by a few loci
  - **polygenic** if it is influenced by a large number of loci with small effects

Mixed models

- A phenotype of interest (x) is a result of major locus, polygenic and environmental effects
- The genotypic effects of AA, Aa and aa can be characterized by
  - z = $-(tq^2 + 2pqdt)$
  - t: $t^2 = 1/[(pq)^2(q + 2pd)^2 + 2pq(d - q^2 - 2pqd)^2 + q^2(1 - q^2 - 2pqd)^2]$
  - q, d = frequency of a, and displacement; they are the free parameters

z	*z+dt*	*z+t*
AA	Aa	aa

- POINTER obtains maximum likelihood estimators by iterating mean and variance of x, d, q, H (polygenic heritability) and B (relative variance due to common environment)

Mixed model

- Write the variance-covariance matrix and likelihood

$$\boldsymbol{\Psi} = 2\mathbf{A}\sigma_{\mathrm{a}}^2 + \mathbf{D}\sigma_{\mathrm{d}}^2 + \mathbf{H}\sigma_{\mathrm{h}}^2 + \mathbf{I}\sigma_{\mathrm{r}}^2$$

$$L(\boldsymbol{\mu}, \sigma_{\mathrm{a}}^2, \sigma_{\mathrm{d}}^2, \sigma_{\mathrm{h}}^2, \sigma_{\mathrm{r}}^2 | \mathbf{Y}) = \frac{1}{(2\pi)^{n/2}|\boldsymbol{\Psi}|^{1/2}} \exp\left[\frac{1}{2}(\mathbf{Y} - \boldsymbol{\mu})'\boldsymbol{\Psi}^{-1}(\mathbf{Y} - \boldsymbol{\mu})\right]$$

$$L(p_{\mathrm{a}}, \mu_{\mathrm{AA}}, \mu_{\mathrm{Aa}}, \mu_{\mathrm{aa}}, \sigma_{\mathrm{a}}^2, \sigma_{\mathrm{d}}^2, \sigma_{\mathrm{h}}^2, \sigma_{\mathrm{r}}^2 | \mathbf{Y}) = \sum_{g}^{G(n)} \tau(g) \frac{1}{(2\pi)^{n/2}|\boldsymbol{\Psi}|^{1/2}} \times \exp\left[-\frac{1}{2}(\mathbf{Y} - \boldsymbol{\mu}_{\mathrm{g}})'\boldsymbol{\Psi}^{-1}(\mathbf{Y} - \boldsymbol{\mu}_{\mathrm{g}})\right]$$

Extension of POINTER

PAP	**POINTER**	Conversion
p frequency of A_1 allele	q frequency of low allele	note if μ_1 is "low"
μ_1 mean of genotype A_1A_1	μ population mean	$\mu = p^2\mu_1 + 2pq\mu_2 + q^2\mu_3$
μ_2 mean of genotype A_1A_2	V population variance	$V = \sigma^2_W + \sigma^2_{ML}$‡
μ_3 mean of genotype A_2A_2	t displacement between homozygotes	$t = \mu_3 - \mu_1$
σ^2_W within genotype variance	d deviation due to dominance	$d = \mu_2 - \left(\frac{\mu_3 - \mu_1}{2}\right)$
H_p proportion of σ^2_W attributed to additive polygenes=(σ^2_A/σ^2_W)	h^2 narrow sense heritability=$\frac{\sigma^2_A}{V}$	$h^2 = H_p\left(\frac{\sigma^2_W}{\sigma^2_W + \sigma^2_{ML}}\right)$

‡Variance due to major locus, $\sigma^2_{ML} = p^2(\mu_1 - \mu)^2 + 2pq(\mu_2 - \mu)^2 + q^2(\mu_3 - \mu)^2$

- Others include FISHER/MENDEL and SAGE (Konigsberg et al. Genet Epidemiol 1989)

Nasopharyngeal carcinoma (Jia et a. 2005)

- A total of 1903 Cantonese pedigrees partitioned into 3737 nuclear families. A mixed as implemented in POINTER shows no evidence for a major gene and the observed data are best explained by a multifactorial mode of inheritance.
- d: where d=0 refers to a recessive major gene, d=0.5 additive and d=1 dominant; t: difference on the liability scale between homozygotes; q: gene frequency; H: the polygenic heritability; Z: the ratio of the multifactorial component in adults and children; and AIC: Akaike's Information Criterion.

Model	*d*	*t*	*q*	*H*	*Z*	*−2 ln L+C*	*AIC*
Sporadic	—	—	—	—	—	−2232.2	−2232.2
Polygenic	—	—	—	0.88	1.00	−4226.8	−4224.8
Multifactorial	—	—	—	0.95	0.61	−4262.2	−4258.2
Recessive	(0)	3.52	0.0694	—	—	−3833.9	−3829.9
Dominant	(1)	2.28	0.0011	—	—	−4144.6	−4140.6
Additive	(0.5)	4.50	0.0013	—	—	−4147.4	−4143.4
Generalized single locus	0.522	4.31	0.0012	—	—	−4147.5	−4141.5

Genetics of Thyroxine in Baboons (Blangero 2005)

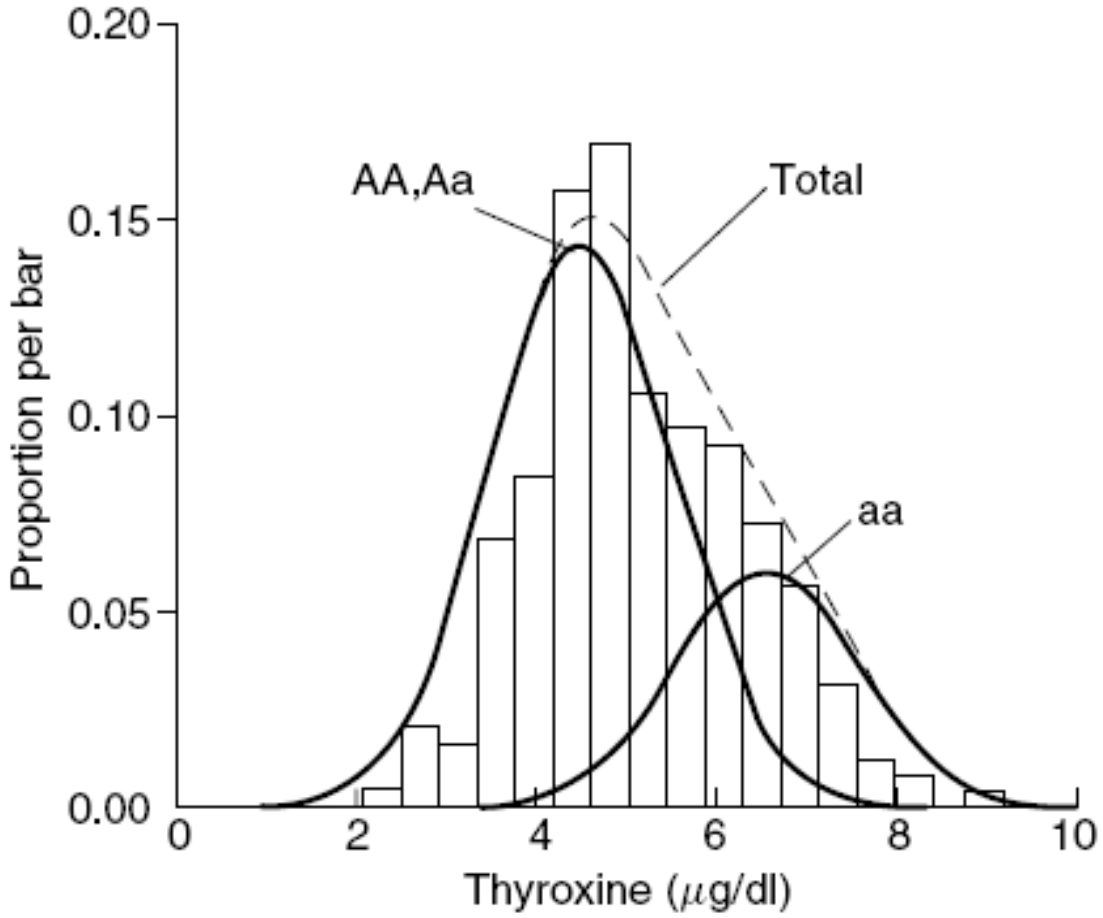
- A major gene influencing thyroid hormone, thryoxine (T_4) as measured by radioimmunoassay in the frozen sera of 248 baboons pedigrees.
- The analysis involved evaluation of a range of models: (1) a general model that allows free transmission probabilities; (2) a finite mixture model with no major locus; (3) τ_{Aa} heterozygote is free; (4) Mendelian recessive model; (5) a polygenic model with not major factor; (6) a sporadic model with no resemblance among relatives.
- It turned out only two component distributions were required, so $\mu_{Aa}=\mu_{AA}$, and best-fitting model was chosen as the one not significantly different from the general model but with minimum AIC.
- All models are significantly worse than the general model but the free τ_{Aa} and recessive models. The latter also has the minimum AIC. There was no evidence for a residual polygenic effect ($h_r^2=0.000$).

Model comparisons

Parameter	General	Environmental	Free τ_{Aa}	Recessive	Polygenic	Sporadic
p_A	0.339	0.460	0.387	0.457	–	–
τ_{AA}	1.000	[0.460]	(1)	(1)	–	–
τ_{Aa}	0.659	[0.460]	0.676	(1/2)	–	–
τ_{aa}	0.153	[0.460]	(0)	(0)	–	–
$\mu_{AA} = \mu_{Aa}$	4.476	4.513	4.420	4.504	5.144	5.141
μ_{aa}	6.592	6.668	6.544	6.580	[5.144]	[5.141]
σ	0.991	0.994	1.017	1.015	1.396	1.395
h_r^2	0.000	0.000	0.000	0.000	0.204	(0)
AIC	16.00	18.81	14.06	13.54	20.16	21.13
Λ	–	8.81	2.06	3.54	14.16	17.13
df	–	3	2	2	4	5
P	–	0.032	0.357	0.316	0.007	0.004

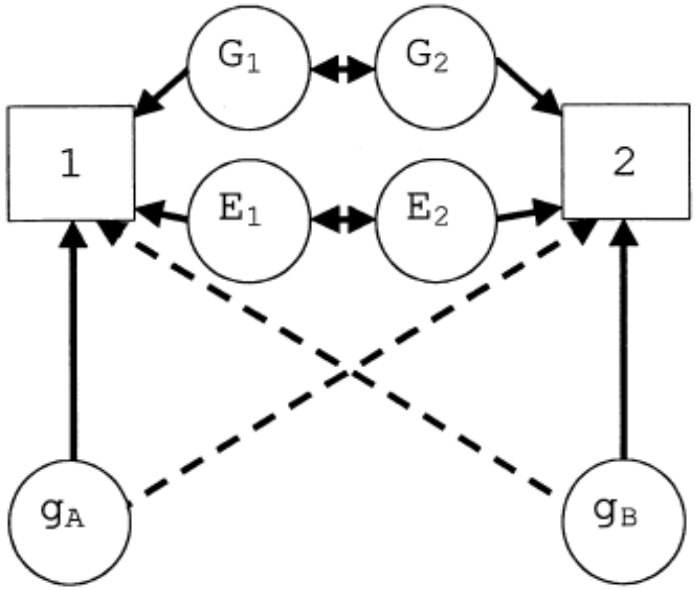
- Λ is the log-likelihood ratio statistic from restricted and general model.

Expected distribution of T_4



The histogram shows the observed distribution. μ_{Aa+AA}=4.5, μ_{aa}=6.58, σ=1.02

Bivariate segregation model



g=major, G=polygenic, E=environment

Study of BMI and fat mass suggested two pleiotropic recessive loci.

Borecki et al. AJHG, 1998

Parameters of the Bivariate Model

Parameter	Description
$p(A)$	Frequency of allele A
$p(B)$	Frequency of allele B
Trait 1:	
m_1	Population mean
$a(A)_1$	Additive genetic effect of locus A
$a(B)_1$	Additive genetic effect of locus B
$d(A)_1$	Dominance effect of locus A
$d(B)_1$	Dominance effect of locus B
SD_1	Residual SD
h_1^2	Proportion of residual variance due to polygenes
Trait 2:	
m_2	Population mean
$a(A)_2$	Additive genetic effect of locus A
$a(B)_2$	Additive genetic effect of locus B
$d(A)_2$	Dominance effect of locus A
$d(B)_2$	Dominance effect of locus B
SD_2	Residual SD
h_2^2	Proportion of residual variance due to polygenes
r_G	Correlation between multifactorial/polygenic components, for traits 1 and 2
r_E	Correlation between nontransmitted environmental components, for traits 1 and 2

Mixed model with genotype-covariate interaction

- Covariates such as age, sex, weight are environmental conditions unique to each individual which may influence the quantitative phenotype differently across the major locus genotypes
- In a study of LDL-C levels in baboons, a comparison of the usual complex segregation analysis and that with genotype-covariate interaction showed support for the latter (Konigsberg et al. Genet Epidemiol,1991).

Regressive models

- Distributions over pedigrees are specified by conditioning each individual's trait value on those of antecedent individuals.
- For continuous trait, they assume (after transformation when appropriate) multivariate normality across members of the individual residuals from the type means.
  - Class A models assume sibling subtypes are dependent only because of common parentage
  - Class D models assume that the sibling correlations are equal but not necessarily due to common parentage alone.
- For binary traits, a multivariate logistic model is formed
- For both continuous and binary traits, finite polygenic mixed model can be used, including binary traits with variable age of onset.

Regressive models (Elston & Anne Spence 2006)

- Assuming a pedigree with n individuals with trait values Y and covariates X.

$$P(\mathbf{Y}|\mathbf{X}) = P(Y_1, Y_2, \ldots, Y_n|\mathbf{X})$$

$$= P(Y_1|\mathbf{X})P(Y_2|Y_1, \mathbf{X})P(Y_3|Y_1, Y_2, \mathbf{X}) \ldots P(Y_n|Y_1, \ldots, Y_{n-1}, \mathbf{X})$$

- Conditioning on information on members of the nuclear family, Y_i is independent of all preceding members. For class D models,

$$P(Y_i|Y_1, \ldots, Y_{i-1}, \mathbf{X}) = P\left(Y_i|Y_{\mathrm{S}}, Y_{\mathrm{F}}, Y_{\mathrm{M}}, \sum_{j<i} Y_j, \mathbf{X}\right)$$

$$P(Y_i|Y_1, \ldots, Y_{i-1}, \mathbf{u}, \mathbf{X}) = P\left(Y_i|Y_{\mathrm{S}}, Y_{\mathrm{F}}, Y_{\mathrm{M}}, \sum_{j<i} Y_j, u_S, u_F, u_M, u_i, \mathbf{u}_j, \mathbf{X}\right)$$

- where u_j specifies the types.

Binary regressive models

- A logistic function can be used for modelling the conditional probabilities with Y_i=0,1, Z_i=2Y_i-1 but 0 if Y_i unobserved and define the logits,

$$\theta_i = \ln\left[\frac{P(Y_i = 1|Y_1, Y_2, \ldots, Y_{i-1}, X_i)}{P(Y_i = 0|Y_1, Y_2, \ldots, Y_{i-1}, X_i)}\right]$$

$$= \beta_{u_i} + \delta_1 Z_1 + \delta_2 Z_2 + \cdots + \delta_{i-1} Z_{i-1} + \xi_i X_i, \quad i = 1, 2, \ldots, n$$

- The likelihood will have the form

$$f(\mathbf{Y}|\mathbf{X}) = \sum_u P(\mathbf{u}) \prod_{i=1}^{n} \frac{e^{\theta_i y_i}}{1 + e^{\theta_i}}$$

Breast cancer

- Claus et al. (1991) showed a dominant major gene model is more preferable to pure environmental, pure polygenic, or recessive major gene models
- Andrieu and Damenais (1997) using Bonney's class D model showed women with a young age at menarche have dramatically higher penetrances than those with older ages at menarche
- This led to the finding of BRCA1 and BRCA2 genes
- A breast and ovarian analysis of disease incidence and carrier estimation algorithm (BOADICEA) is available http://www.srl.cam.ac.uk/genepi/boadicea/boadicea_home.html as with R/Bayesmendel, http://www.cancerbiostats.onc.jhmi.edu/BayesMendel

(Thomas (2004) *Statistical Methods in Genetic Epidemiology* and Blangero J (2005) in Encyclopedia of Biostatistics II)

Generalized estimating equations (GEE)

- Recall that generalized linear models (GLM) is an extension of linear regression from normal to exponential family

$$f_{Y_i}(y_i;\theta_i,\phi_0)=\exp\left\{\frac{\theta_i y_i - b(\theta_i)}{a_i(\phi_0)}+c_i(y_i,\phi_0)\right\}$$

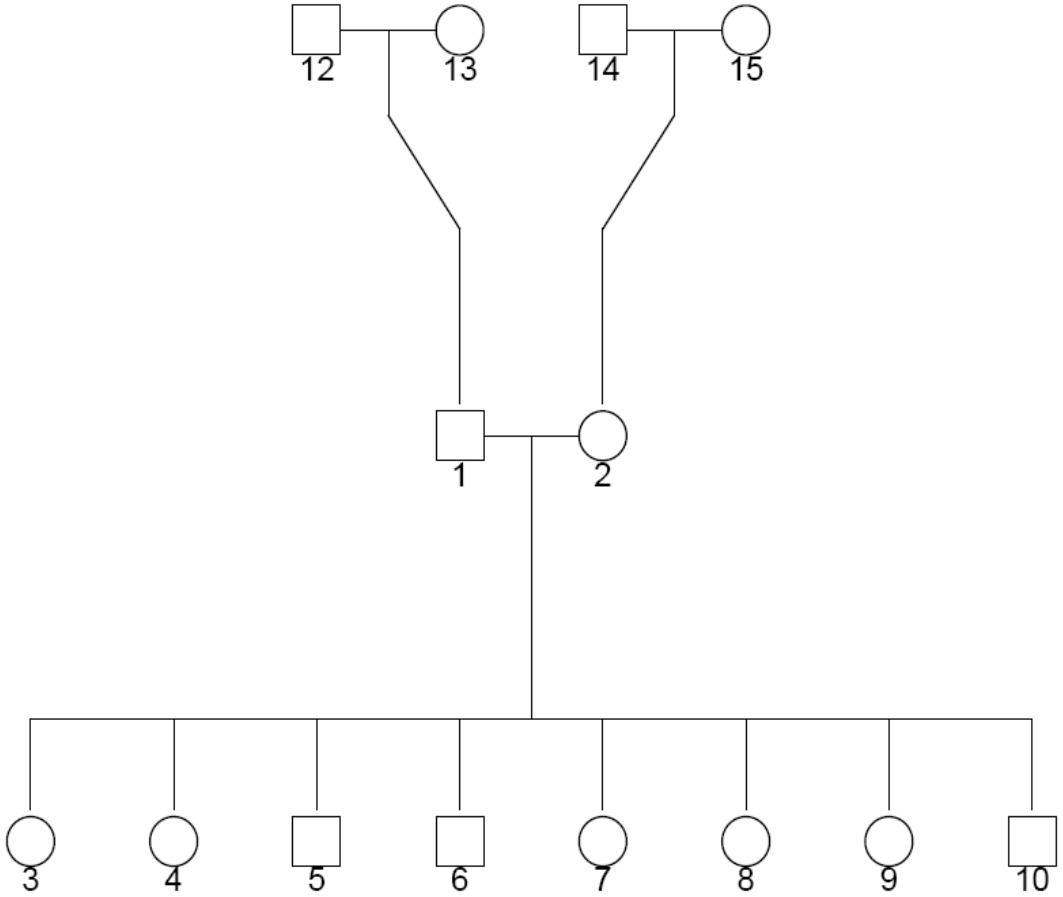
- has been extended to quasi-likelihood through estimating equations of the form

$$U_j(\boldsymbol{\beta})=\sum_{i=1}^{n}\frac{(y_i-\mu_i)}{\phi w_i^{-1}v(\mu_i)}\frac{\partial\mu_i}{\partial\beta_j}=0,\quad j=1,2,\ldots,p$$

- and further to GEE by simultaneously estimating parameters in the covariance matrix of the response vector. Method is available for weighting the estimating equations to the case of missing at random (MAR) and for multistage designs.

Gene expression levels and aging (Tan et al. 2008)

Microarray gene expression data in 194 individuals from 14 CEPH Utah families were obtained by hybridising RNA extracted from immortalised lymphoblastoid cells to the Affymetrix focus array containing ~8500 genes, the analysis focused on demographic characteristics such as age and sex on differential gene expression patterns.



Analysis

- Normalisation was performed for pre-preprocessing data.
- The major statistical method is GEE, which accounts for within individual correlation. In this technique one has the flexibility to specify correlation structures. It has been implemented in R (as well as Stata, SAS). To simplify the analysis, the parent generation is dropped from the analysis. The sibling correlation is also quantified with intraclass correlation (ICC), empirical significance is assessed by resampling method implemented in R.
- To account for multiple-testing, the false-discovery rate (FDR) is used.
- Hierarchical cluster analysis is used for the significant genes with *gplot* package.
- EASE software (http://david.abcc.ncifcrf.gov/ease/ease.jsp) is used to cluster significantly regulated genes into pathways.
- The Affymetrix Focus database was used to extract information on gene annotation.

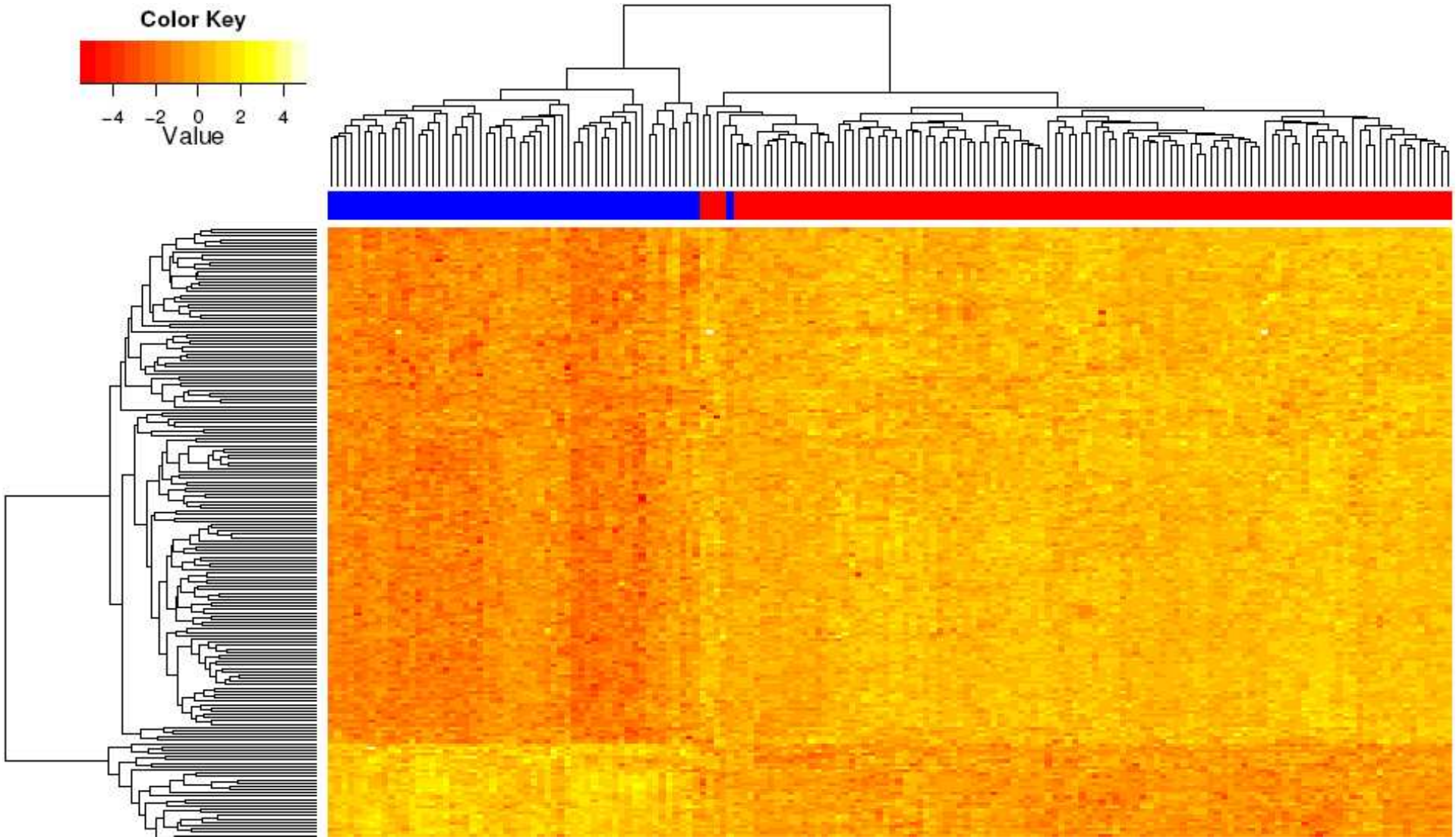
200 topmost significant genes

Gene Category	List hits	Pop. hits	score
Cell-cell signaling	30	495	0.000
Cell communication	68	2402	0.008
Inorganic anion transport	6	60	0.009
Channel/pore class transporter activity	14	290	0.010
Chloride transport	5	42	0.012
Signal transducer activity	51	1733	0.014
Voltage-gated ion channel activity	8	119	0.014
Alpha-type channel activity	13	280	0.018
Anion transport	7	98	0.019
Extracellular	32	979	0.022
Ion channel activity	12	258	0.024
Ion transport	17	433	0.025
Muscle contraction	8	136	0.028
Development	42	1422	0.028
Receptor binding	17	441	0.028
Cell surface receptor linked signal transduction	28	875	0.036
Sex differentiation	3	14	0.036
Organogenesis	26	804	0.039
Voltage-gated chloride channel activity	3	15	0.040
Transcription factor complex	17	458	0.043
Enzyme linked receptor protein signaling pathway	8	151	0.045
TGFbeta receptor signaling pathway	4	38	0.048

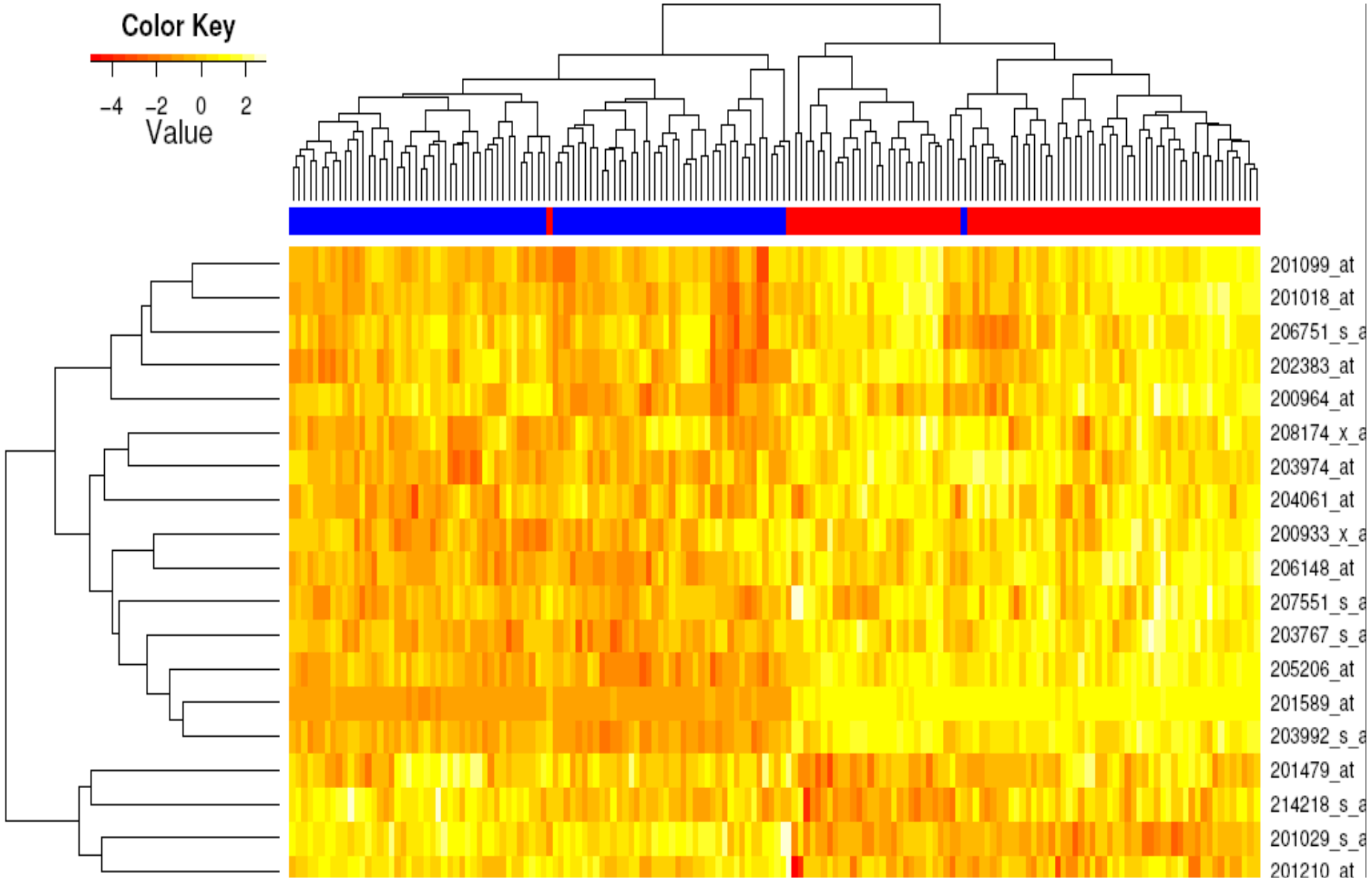
101 significant sex-regulated genes in the young

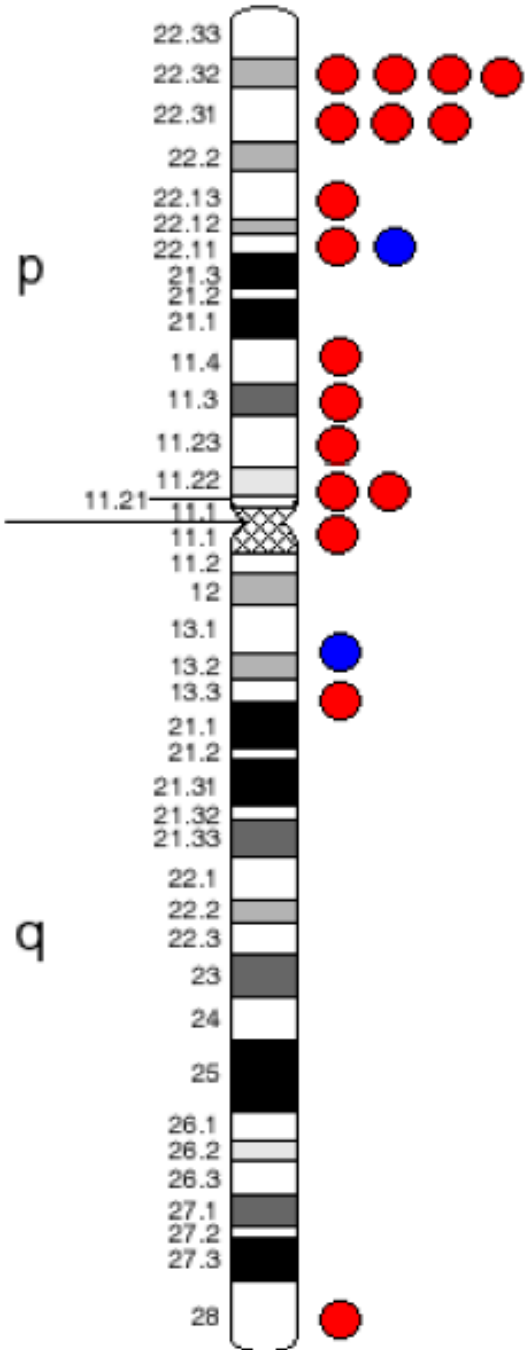
System	Gene Category	List hits	Population hits	EASE score
Molecular Function	RNA binding	15	375	0.000
Molecular Function	Nucleic acid binding	35	1897	0.002
Cellular Component	Ribonucleoprotein complex	10	290	0.006
Biological Process	RNA processing	9	248	0.007
Cellular Component	Spliceosome complex	5	69	0.008
Biological Process	RNA metabolism	9	269	0.012
Cellular Component	cAMP-dependent protein kinase complex	3	15	0.013
Cellular Component	Obsolete cellular component	10	324	0.013
Biological Process	RNA splicing	5	92	0.021
Biological Process	Development	25	1422	0.027
Biological Process	Male gamete generation	5	102	0.029
Biological Process	Spermatogenesis	5	102	0.029
Biological Process	Sexual reproduction	6	152	0.030
Molecular Function	Pre-mRNA splicing factor activity	4	61	0.030
Biological Process	Reproduction	6	153	0.030
Cellular Component	Nucleus	34	2101	0.032
Molecular Function	Cell adhesion molecule activity	8	277	0.035
Molecular Function	Chromatin binding	3	29	0.041

200 topmost significant genes regulated by age (Nearly all subjects are clearly distinguished left=grandparents, right=grandchildren)



The top 19 X-linked genes (Nearly all young males are to the left and females to the right panel)





Location of the 19 sex regulated significant genes (marked as red circles and pseudoautosomal genes marked blue) on the X-chromosome in the young group. There is an obvious concentration on the short arm (Xp) especially to the extreme end of Xp

Summary

- Analysis of gene expression data in the CEPH Utah families has identified interesting patterns in gene expression regulation. Differential expression analysis across age groups identified the decreased cell-cell signaling as an important biological process involving human aging. While sex-dependent gene expression is predominated by genes escaping X-inactivation, such a pattern is not affected by the aging process.
- Sibship correlation analysis suggests that genes whose transcriptomic activities are under genetic control are those relatively active genes but not vice versa.

Software

- Familial aggregation
  - FISHER/SOLAR, PATHMIX, LISREL, Mx, Mplus
  - Recast in gllamm framework so that they can be modelled through Stata and S-PLUS/R (Rabe-Hesketh et al. (2008) Biometrics 64:280-8)
- Segregation analysis
  - POINTER, MENDEL, PAP, SAGE
- GEE
  - SAS, Stata, R

To wrap up

- To establish familial aggregation is traditionally the first step in genetic study of complex traits.
- Segregation analysis attempts to infer mode of inheritance through analysis of family data. It is most commonly conducted in likelihood framework. Complex segregation analysis considers multiple mating types and multiple determinants.
- It involves three types of parameters: allele frequencies, penetrance (genotypic or ousiotypic means), and transmission probabilities.
- Advantage and disadvantages: best-fitting model is not necessarily correct AND inclusion of genetic markers increases the power of genetic analysis, nevertheless it is fundamental to the understanding of linkage analysis.